United Nations ECE/CES/GE.41/2017/11

Economic and Social Council

Distr.: General
14 July 2017

Original: English
English and Russian only

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Nineteenth Meeting
Geneva, 4–6 October 2017
Item 4 of the provisional agenda
Evaluating the census and measuring data quality

Quality measurement in the state information systems and databases

Note by Statistics Estonia[1]

Summary

The arrangement of the first register-based census in Estonia requires a relatively high volume of preparatory work for improvement of data quality, advancement of semantical capacity etc.

A new national level initiative, which maps institutional knowledge and skills necessary for creating a result satisfying certain functional requirements while arranging a register-based census in Estonia in 2021. Estonian State Information System Authority (RIA) ordered a study in 2016, which aim was to create the manual of data quality measurement for the register holders.

Data quality in registers databases directly affects the functioning of the state. The better the quality in databases, the more comfortable and more easily accessible services are available to the general public and more accurate decisions can be made for statistical purposes.

This document is dedicated to introduce a new initiative at national level.

[1] Prepared by Kristi Lehto and Diana Beltadze

GE.17-11842(E)
1711842

Contents

I. Introduction

1. A register-based population and housing census (REGREL) requires availability of complete, regularly updated and verifiable state registers that cover the main questions of the census.

2. From 2009, Statistics Estonia has made preparations for conducting the 2021 population and housing census in a register-based format. The register-based census is supported by a favourable context: nearly all mandatory EU census characteristics (EU Regulation No 763/2008) can be extracted from Estonian registers; there is a functioning system of personal identification codes, enabling identification in all personal data registers; and the Address Data System (ADS), created and launched in 2008, is being integrated in many databases.

3. In 2015, Statistics Estonia issued requirements for register holders, which they have to meet before the start of the register-based census:

- At least 97% coverage of the population;
- Information on required census characteristics must be regularly updated and this process must be documented;
- At least 95% of entries must have an identifier in the standard format;
- At least 95% coverage of relevant census characteristics;
- The rate of significant or technical errors in the values of relevant census characteristics may not exceed 1%.

4. The measures required to ensure compliance of all registers with the requirements include: improvement and monitoring of data quality; replacing other address data sources with the administrative system of the ADS; providing data with classifications. 5. Harmonisation of classifications and terminology in state registers is one of the key activities in the preparation of the register-based census.

5. Pursuant to §50 (2) of the Official Statistics Act, Statistics Estonia shall assess the quality of data in databases. Requirements for ensuring data quality in the census:

- Registers must follow up on the results of regular quality assessment conducted by Statistics Estonia;
- Data acquired from registers must be compliant with the characteristics as agreed with Statistics Estonia, they must be complete and accurate;
- Registers must verify internal consistency of their data (e.g., to detect and correct duplicate entries and errors, which are identifiable by combining different characteristics, etc).

6. 7. The main focus in the preparations for the register-based census is, since 2015, on the content and quality of data:

- For one fifth of the population, the registered place of residence is different from the actual place of residence;
- The quality of data in the State Register of Construction Works is insufficient for statistical purposes. The low quality of the Register of Construction Works is caused by under-coverage of buildings and dwellings, incomplete data on technical characteristics, and lack of updates.

II. Is the quality of register data only a concern for statisticians?

7. The implementation of REGREL is based on the provisions of §50 of the Official Statistics Act (OSA). According to this, Statistics Estonia shall assess the quality of the basic data of databases and make proposals to the chief processor of a database for improving the quality of data. There are no irresolvable legal problems in this context, unless we consider the fact that the producer of official statistics is charged with the task of supervising the quality of data in databases for statistical purposes, but the extent of this task has not been specified. It is unclear who should be responsible for drawing up quality standards for state registers.

8. It also leads to questions about the format of feedback on data quality to be provided to registers. There are no current standards to regulate this.

9. As an example, we can use the database of dwellings. In order to improve the quality of data, the registrar of the Estonian Register of Buildings has to undertake work, which requires clear legal regulation. For instance, if the owner of a building has submitted data to the register based on a legal document at the beginning of the previous century, but the current owner is a resident of another country, there is no obligation for the owner to verify and update the data on his property in the Estonian Register of Buildings. As a result, the database includes a number of data entries that do not conform to current data quality requirements, which in turn creates problems for the production of statistics, even though the data are legally valid.

10. The other side of the coin is that databases do not have sufficient resources and the work required to improve data quality is not integrated in their daily operations.

11. In a 2016 meeting of the working group on REGREL databases, many registrars raised the question: How to improve the quality of data and what other benefits would it offer beyond the production of statistics? The universal response of Statistics Estonia was that better data facilitate better decisions.

12. High-quality decisions can only be based on verified databases of high quality. Verification of data quality is not productive if it remains a one-time effort. Monitoring of data quality must be a comprehensive and continuous process. It requires planning, measurement, analyses, tests and, if necessary, (legal) measures for improving the data.

13. A universal data quality management model was created to improve the general situation in the country.

III. Data quality manual for registers

14. Estonian State Information System Authority (RIA) ordered a study in 2016 with the aim to create a manual of data quality measurement for the register holders. The study was conducted by AS PricewaterhouseCoopers Advisors.

15. RIA’s goal was to develop a comprehensive, simple and implementable methodology for assessing the current state, identifying the optimal required quality level, planning developments for individual levels, and creating procedures for moving from one level to the next. The work on the development of the data quality manual resulted in the specification of database use and administration methodologies, techniques, organisational structures, roles and key functions, which facilitate technical and semantic improvements in the quality of register data.

16. The developed data quality manual for the state includes methodologies for measuring and ensuring the quality of data in the state information system as a whole. The manual specifies a methodology for managing the monitoring and supervision of database quality and includes recommendations for metrics to be used in data quality monitoring.

17. The developed framework for data quality management includes three elements:

- Data quality maturity model, developed for measuring and improving the categories associated with data quality management. The developed model covers five management categories at five levels of maturity. The description of the model also includes implementation methodology;
- Set of data quality indicators, which can be used for testing different aspects of data quality. The set includes nine indicators, each provided with data quality requirements, measures to ensure quality, and test questions;
- Framework for data quality management, which is a set of iteratively implementable actions to ensure data quality, based on the OPDCA quality management model.

18. The framework integrates the listed elements into a single quality management system.

19. The aim of the study of data quality indicators was to define a universal set of data quality indicators, which could be adopted by all databases of the state information system for measuring and ensuring quality.

20. The sources of indicators were identified in a web research to find in the public information environment descriptions of hypothetical or implemented data quality frameworks, which include defined indicators for measuring data quality.

21. The research resulted in a review of about 20 descriptions of data quality frameworks, including data quality standards, data quality implementation practices in international banks and enterprises, data quality management practices in public authorities of different countries, articles published by experts of the field, and blog posts in widely recognised online research and IT publications.

22. The screening of indicators resulted in a set of nine indicators, which were found to be suitable for assessing the quality of data in the databases of the state information system. A description of selected indicators is provided in Table 1.

Table 1
Set of data quality indicators

Accuracy	Data accuracy refers to the accurate representation of the true values of a term or event by data attributes in a specific context of use.
Completeness	Data completeness refers to the availability of entered values for all data attributes of all entities where values are required.
Consistency	Data consistency refers to the lack of discrepancy between data and consistency with other data.
Credibility	Data credibility refers to data attributes, which can be considered as true and reliable by users in a specific context of use.
Timeliness	Data timeliness refers to the adequacy of temporal distance between data record and the underlying event.

Confidentiality	Data confidentiality means availability of data to authorised users (persons or technical systems) only and unavailability to everyone else.
Singularity	Data singularity refers to compliance with the applicable principles of singular data collection from providers.
Non-redundancy	Non-redundancy refers to planning and development measures to prevent redundancy in data structures of the database and measures to prevent redundancy in data processing.
Regularity	Data regularity refers to existence of data attributes, which conform to data quality standards, conventions, applicable legislation or equivalent norms in a specific context of use.

23. The aim of the study of data quality management was to create a methodology of data quality measurement for the databases of the state information system. The study task was to analyse the practices adopted by organisations for measuring data quality processes and to develop a methodology for measuring the quality of data in the databases of the state information system.

24. Data quality aspects were grouped in five main categories of data quality assurance:

- Management and planning. This category includes aspects of activities and tools for comprehensive management and planning of data quality, incl. management of rules, policies, statutes and plans, strategic approach to data quality adopted by the organisation;
- Organisation and responsibilities. This category covers the aspects associated with the establishment of an organisation responsible for managing data quality and with assignment of responsibilities;
- Processes. This category deals with implementation of data quality plans, incl. monitoring, measurement, assurance;
- Knowledge and competencies. This category includes aspects related to the management of knowledge on data quality (incl. knowledge dissemination and engagement of competencies);
- Technical tools. This category covers the aspects associated with adoption and development of IT tools to support data quality processes.

25. The next step was to identify a suitable methodology for assessing the data quality situation in each individual category. According to the raised hypothesis, a field-specific maturity model could be used as such a tool, because the maturity model is a common tool used for assessing people/culture, processes/structures, and objects/technologies. The maturity model is a tool that facilitates scale-based assessment of the current situation in a particular field and establishment of gradual targets for development.

26. The chosen maturity model was based on the following five levels:

(a) Reactive level. Data quality processes are unpredictably, weakly controlled, and actions on processes are reactive in nature;

(b) Controlled level. The importance of data quality management has been acknowledged and repeatable procedures have been introduced;

(c) Standardised level. Data quality processes have been standardised and the quality of data has been checked for conformity;

(d) Managed level. Processes have become sustainable. The results of data quality measurement are used for managing existing processes;

(e) Optimised level. Data quality processes are reviewed and updated on a regular basis.

27. The study resulted in the development of a data quality maturity model for assessing the maturity of each data quality category based on five maturity levels. The model enables adopters to measure the maturity of data quality management on a 5-point scale (see Annex).

IV. Results of the pilot study

28. The completion of the data quality manual for register holders was followed by a study of data quality of the databases to test the applicability of the manual and to supplement the manual with the results of feedback from practical application.

29. The study of applicability of the data quality manual, commissioned by RIA, was conducted in three state registers: Population Register (PR), Address Data System (ADS), and Administration system for the state information system (RIHA).

30. The maturity of data quality assurance was assessed using the maturity assessment tool (see Annex). Assessments were obtained by comparing the statements describing the maturity levels in different categories with the actual situation.

31. The assessment was conducted in two rounds. In the first round, some statements in the maturity model remained incomprehensible for participants, after which the assessment tool was modified to increase comprehensibility, and then a second round of assessment was carried out. For the Population Register, the first and second round resulted in different assessments of maturity level.

32. For each register, the assessment of maturity levels resulted in the determination of:

- Maturity levels by categories, indicating the maturity of data quality management in each category included in the maturity model (scale 1–5);
- Aggregate maturity rating, which can be used for detailed measurement of the efficiency of actions to improve data quality.

Table 2
Data quality assurance maturity levels

Maturity category	*RIHA*	*PR (1)*	*PR (2)*	*ADS*
1. Management and planning	1	1	3	5
2.Organisation and responsibilities	1	3	3	5
3.Processes	1	1	4	5
4.Knowledge and competencies	1	2	5	5
5.Technical tools	1	5	5	5
Aggregate maturity rating	**1.0**	**2.4**	**4.0**	**5.0**

33. The following general recommendations were issued for implementation of the data quality manual and for quality assessment:

- In case of the databases of the state information system, always observe data quality by individual databases and include all data;
- When zoning an organisation, consider including all stakeholders that affect data quality (incl. authorised database processors and IT specialists) in the zone of data quality management;
- When assessing applicability of statements that describe maturity levels, immediately note for each statement the reasons why the respective assessment was made. This completes a major step towards subsequent compilation of an improvement plan. It also helps to reduce the time required for developing the plan and to prevent subsequent inconsistencies with original assessments;
- Include the organisation's IT specialists in the assessment to provide a competent assessment of the state of technical tools.

34. Among other things the features of Estonian National Metadata Catalogue were worked out for register data holders: comprehensive, mandatory, serves as a base register for a number of processes. New vision includes: distributed metadata production and publication; publication in machine-readable format; harvesting into central repository; opening up agencies internal metadata repositories; more realistic approach to metadata production requirements.

V. Conclusions

35. The aim of the study was to develop the methodology for measuring the data quality and to create the structured manual for the implementation activities. RIA also carried out the pilot study and evaluated the data quality in Population Register (PR), Address Data System (ADS) and Administration system for the state information system (RIHA).

36. Key conclusions for the use of the data quality manual:

- Recognition of the importance of data quality by the management;
- Creation of a continually operating sustainable system;
- Clear division of responsibilities and coverage assurance.

37. Input data quality is very important to every register-based census country. It is essential that data quality is a priority to the register holders.

38. Data Quality Initiative for improving data quality in registers is based on well-known data quality models (quality attributes) and process improvement frameworks (5-level maturity model) there was developed data quality guidance document and carried out a survey in 2016 of data quality in three Estonian registers.

References

Data quality manual. (2016) https://www.ria.ee/ee/andmekvaliteedi-tagamine.html (available only in Estonian).

EU Regulation No 763/2008
http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32008R0763.

Official Statistics Act https://www.riigiteataja.ee/en/eli/ee/506012015002/consolide.

Annex

Questionnaire to determine the maturity level of data quality assurance

DATA QUALITY MANAGEMENT AND PLANNING

1 Data quality requirements for the database have been specified and documented (1, 2).

2 Data quality has been checked for conformity to requirements (3).

3 All critical data, which are subject to policies, have been specified and documented (2).

4 Data quality policies have been documented and disclosed (3).

5 Data quality management has been implemented in a uniform manner in the zone concerned (4).

6 Data quality is measured and improved based on the analysis of measurement results (4).

7 Data quality indicators are linked to higher-level strategies or policies (4).

8 Data quality processes are subject to continuous improvement (5).

9 Data quality is managed through regular review, amendment and disclosure of policies (5).

MANAGEMENT OF DATA QUALITY RESPONSIBILITIES IN THE ORGANISATION

10 The data quality process has an identifiable owner (1, 2).

11 The responsibilities of the owner of data quality processes are clearly defined in job description (3).

12 The owners of data have been specified (1, 2, 3).

13 A senior management group has been established to take responsibility for data quality (4).

14 The group includes representatives of other stakeholders of the respective register (4).

15 The group regularly reviews and updates the responsibilities for data quality management (5).

DATA QUALITY PROCESSES

16 A data quality profile has been specified and it is used to detect weaknesses in the process (1, 2).

17 Measurable data quality indicators have been specified and documented (1, 2).

18 Data quality is measured and weaknesses are detected at an early stage of the process (3).

19 The organisation looks for ways to prevent problems (1, 2).

20 All detected issues are registered and the resolution process is traceable (1, 2).

21 Data quality assurance processes have been specified and documented (1, 2).

22 The impact of weaknesses in data quality has been determined (2).

23 Data quality measurements are conducted on a regular basis (4).

24 The measurement results are available to the management (4).

25 An action plan for data quality improvement has been specified and documented (4).

26 Investigation of root causes of data quality weaknesses is a common practice (4).

27 Data quality indicators are regularly reviewed to identify further opportunities for process improvement (5).

KNOWLEDGE AND COMPETENCIES ON DATA QUALITY

28 Training on data quality has been provided to raise awareness of data quality issues (1, 2).

29 Key persons make plans and give recommendations on data quality management (1, 2).

30 Employees who can affect data quality have completed data quality training (3).

31 Unofficial mentorship is provided to raise awareness of data quality issues (3).

32 Best practices for data quality assurance are documented as a shared knowledge base (3).

33 Data quality training is organised on a regular basis (4).

34 The content of data quality training is regularly reviewed and updated as necessary (5).

TECHNOLOGICAL TOOLS FOR DATA QUALITY ASSURANCE

35 Data quality standards for IT tools have been developed and documented (1, 2).

36 Efforts have been made to create databases conforming to the ‘single source of truth’ model (2).

37 Tools for measuring and improving the quality of data have been introduced and are being used (1, 2, 3).

38 Conformity to data quality standards is monitored at the stage of new project proposals (3).

39 Data quality reporting tools include analytics features to assist in management decisions concerning data quality (4).

40 A target portfolio of development actions for IT systems that support data quality management has been agreed (4).

41	Automated data correction or verification procedures have been implemented based on the data quality profile (4).
42	Primary data principles have been adopted in the development of databases (5).
43	The target portfolio of development actions for IT systems that support data quality management is regularly reviewed and updated as necessary (5).